Khan *et al. Journal of Cloud Computing: Advances, Systems and Applications* (2015) 4:2
DOI 10.1186/s13677-015-0026-8

Journal of Cloud Computing
a SpringerOpen Journal

RESEARCH Open Access

Towards cloud based big data analytics for smart future cities

Zaheer Khan[1*], Ashiq Anjum[2], Kamran Soomro[1] and Muhammad Atif Tahir[3]

Abstract

A large amount of land-use, environment, socio-economic, energy and transport data is generated in cities. An integrated perspective of managing and analysing such big data can answer a number of science, policy, planning, governance and business questions and support decision making in enabling a smarter environment. This paper presents a theoretical and experimental perspective on the smart cities focused big data management and analysis by proposing a cloud-based analytics service. A prototype has been designed and developed to demonstrate the effectiveness of the analytics service for big data analysis. The prototype has been implemented using Hadoop and Spark and the results are compared. The service analyses the Bristol Open data by identifying correlations between selected urban environment indicators. Experiments are performed using Hadoop and Spark and results are presented in this paper. The data pertaining to quality of life mainly crime and safety & economy and employment was analysed from the data catalogue to measure the indicators spread over years to assess positive and negative trends.

Keywords: Big data, Data mining and analytics, Smart city, Cloud computing

Introduction

ICT is becoming increasingly pervasive to urban environments and providing the necessary basis for sustainability and resilience of the smart future cities. With the rapid increase in the presence of Internet of Things (IoT) and future internet [1,2] technologies in the smart cities context [3-5], a large amount of data (a.k.a. big data) is generated, which needs to be properly managed and analysed for various applications using a structured and integrated ICT approach. Often ICT tools for a smart city deal with different application domains such as land use, transport and energy, and rarely provide an integrated information perspective to deal with sustainability and socioeconomic growth of the city. Smart cities can benefit from such information using big, and often real-time, cross-thematic data collection, processing, integration and sharing through inter-operable services deployed in a cloud environment. However, such information utilisation requires appropriate software tools, services and technologies to collect, store, analyse and visualise large amounts of data from the city environment, citizens and various departments and agencies at city scale to generate new knowledge and support decision making.

The real value of such data is gained by new knowledge acquired by performing data analytics using various data mining, machine learning or statistical methods. However, the field of smart city based data analytics is quite broad, complex and is rapidly evolving. The complexity in the smart city data analytics manifests due to a variety of issues: i) Requirements of cross-thematic applications e.g. energy, transport, water, urban etc, and ii) multiple sources of data providing unstructured, semi-structured or structured data, and iii) trustworthiness of data [6,7]. In this regard, this paper provides a data oriented overview of smart cities and provides a cloud based analytical service architecture and implementation for the analysis of selected case study data.

Smart cities provide a new application domain for big data analytics and relatively not much work is reported in literature. A review of the state of the art provides very promising insights about applying cloud computing resources for large scale smart city data analytics. For instance, Lu et al. [8] focus on using computational resources for large scale data for climate having complex

*Correspondence: Zaheer2.Khan@uwe.ac.uk
[1]Faculty of Environment and Technology, Department of Computer Science and Creative Technologies, University of the West of England, Bristol, UK
Full list of author information is available at the end of the article

structure and format. Using a multi scale dataset for climate data, they demonstrated a cloud based large scale data integration and analytics approach where they made use of tools such as RapidMiner and Hadoop to process the data in a hybrid cloud. Among others, the COSMOS project [9] provides a distributed on-demand cloud infrastructure based on Hadoop for analysing Big Data from social media sources. The infrastructure has the capability to process millions of data-points that would take much longer on a desktop computer. It allows social scientists to integrate and analyse data from multiple non-interoperable sources in a transparent fashion. Such a Big Data analysis platform can also be useful for smart cities as it would allow decision-makers to collect and analyse data from many sources in a timely manner. Ahuja and Moore [10] provide a state of the art review of the technologies being used for big data storage, transfer and analysis. Qin et al. [11] present challenges of Big data analytics and acknowledge the capabilities of MapReduce and RDBMS to solve these challenges. The main contribution of their work is that they have provided a unified MapReduce and RDBMS based analytic ecosystem to avail complementary advantages from both systems. Recently some studies have investigated the usefulness of data mining techniques to combine data from multiple sources such as by Moraru and Mladenic [12]. They applied Apriori technique, which is rule based data mining technique, to learn rules from data. Although they are able to extract some rules from small scale but they're unable to learn much on large scale data due to high volume of the data and the limited memory on a single system.

We use a similar approach that is based on MapReduce. Our prototype implementation analyses the Bristol open dataset to identify correlations between selected urban environment indicators such as Quality of Life. We have developed two implementations using Hadoop and Spark to compare the suitability of such infrastructures for Bristol open data analysis.

The remainder of this paper is structured as follows: the next section provides background and rationale in the context of smart cities. Section "An abstract architectural design of the cloud-based big data analysis" provides a data analytics service architecture and design for analytical processing of big data for smart cities. After this, a simple use case based on Bristol open data by identifying needs of information processing and knowledge generation for decision making is presented in Section "A use case: analytics using Bristol open data". In Section "Prototype implementation" we present the applicability of the proposed solution by implementing a MapReduce based prototype for Bristol open data and discuss outcomes. Finally, we conclude our discussion and present future research directions in Section "Conclusions and future directions".

ICT and smart cities

Approximately 50% of world's population live in urban areas, a number which is expected to increase to nearly 60% by 2030 [13]. High levels of urbanisation are even more evident in Europe where today over 70% of Europeans live in urban areas, with projections that this will increase to nearly 80% by 2030 [13,14]. A continuous increase in urban population strains the limited resources of a city, affects its resilience to the increasing demands on resources and urban governance faces ever increasing challenges. Furthermore, sustainable urban development, economic growth and management of natural resources such as energy and water require better planning and collaborative decision making at the local level. In this regard, the innovation in ICT can provide integrated information intelligence for better urban management and governance, sustainable socioeconomic growth and policy development using participatory processes [15].

Smart cities [4] use a variety of ICT solutions to deal with real life urban challenges. Some of these challenges include environmental sustainability, socioeconomic innovation, participatory governance, better public services, planning and collaborative decision-making. In addition to creating a sustainable futuristic smart infrastructure, overcoming these challenges can empower the citizens in terms of having a personal stake in the well-being and betterment of their civic life. Consequently, city administrations can get new information and knowledge that is hidden in large-scale data to provide better urban governance and management by applying these ICT solutions. Such ICT enabled solutions thus enable efficient transport planning, better water management, improved waste management, new energy efficiency strategies, new constructions and structural methods for health of buildings and effective environment and risk management policies for the citizens. Moreover, other important aspects of the urban life such as public security, air quality and pollution, public health, urban sprawl and bio-diversity loss can also benefit from these ICT solutions. ICT as prime enabler for smart cities transforms application specific data into useful information and knowledge that can help in city planning and decision-making. From the ICT perspective, the possibility of realisation of smart cities is being enabled by smarter hardware and software e.g. IoTs i.e. RFIDs, smart phones, sensor nets, smart household appliances, and capacity to manage and process large scale data using cloud computing without compromising data security and citizens privacy [16]. With the passage of time, the volume of data generated from these IoTs is bound to increase exponentially and classified as Big data [17]. In addition, cities already possess land use, transport, census and environmental monitoring data which is collected from various local, often not interconnected, sources and used by application specific systems

but is rarely used as collective source of information (i.e. system of systems [18]) for urban governance and planning decisions. Many local governments are making such data available for public use as "open data" [19]. Managing such large amount of data and analysing for various applications e.g. future city models, visualisation, simulations, provision of quality public services and information to citizens and decision making becomes challenging without developing and applying appropriate tools and techniques.

In the above context, recent emergence of Cloud computing promises solutions to such challenges by facilitating big data storage and delivering the capacity to process, visualise and analyse city data for information and knowledge generation. Such a solution can also facilitate the decision makers in meeting the QoS requirements by providing an integrated information processing and analytic infrastructure for variety of smart cities applications to support decision-making for urban governance.

Figure 1 depicts our view of the main thematic pillars of smart cities: smart people, smart economy, smart environment, smart governance and smart mobility which contribute towards the sustainability of resources and resilience against increasing urban demands. The main motive towards developing such a view is to consider a holistic approach for smart cities by providing data acquisition, integration, processing and analysis mechanisms to synthesise the needed information that can help in enhancing resilience and sustainability of a city. Managing data for these thematic domains in a Cloud environment provides the opportunity to integrate data acquired from various sources, process and analyse it in acceptable timeframes. However, it is not straightforward to adopt cloud computing to deal with smart city applications due to a number of challenges and requirements [20]. Our aim here is to discuss a perspective on how these challenges can be addressed in part by using ICT tools and software services to intelligently manage and analyse the complex big data of smart cities, by incorporating a suitable Cloud architecture [4,15,21].

An abstract architectural design of the cloud-based big data analysis

This section discusses the development of a cloud service for smart city related big data analysis. Firstly we describe the design and implementation of a generic Cloud based Analytics Service. We then discuss the process used to exploit this service for analysing the Bristol Open data. Our guiding design principle for the Cloud-based analytic service is to reuse existing, well-tested tools and techniques.

The system architecture, as shown in Figure 2, is divided into three tiers to enable the development of a unified knowledge base. Each layer represents the potential

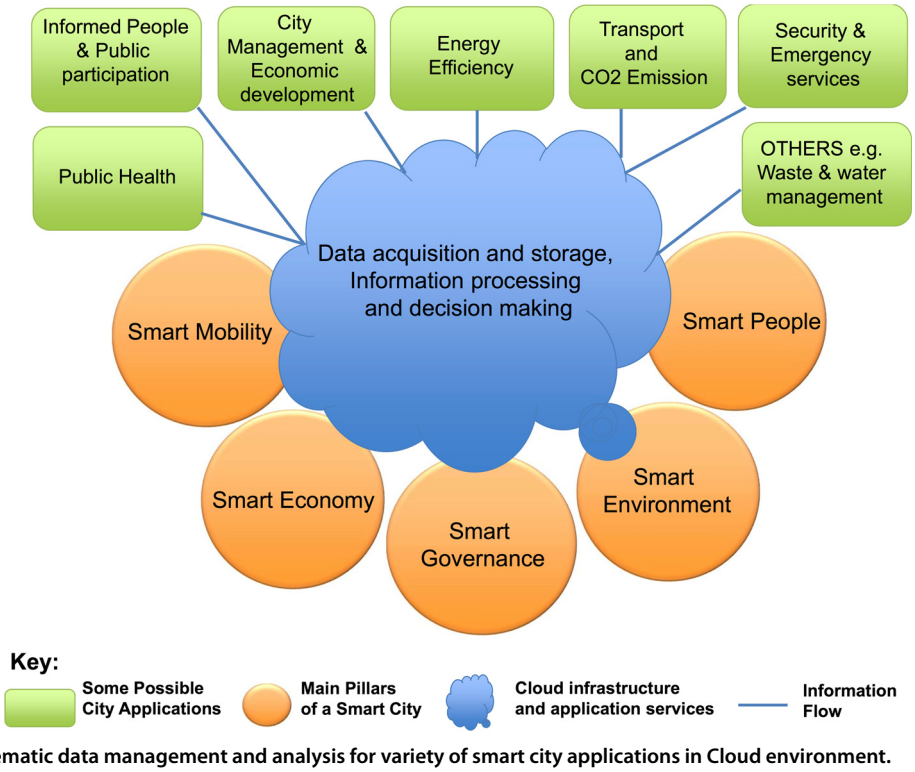

Figure 1 Cross-thematic data management and analysis for variety of smart city applications in Cloud environment.

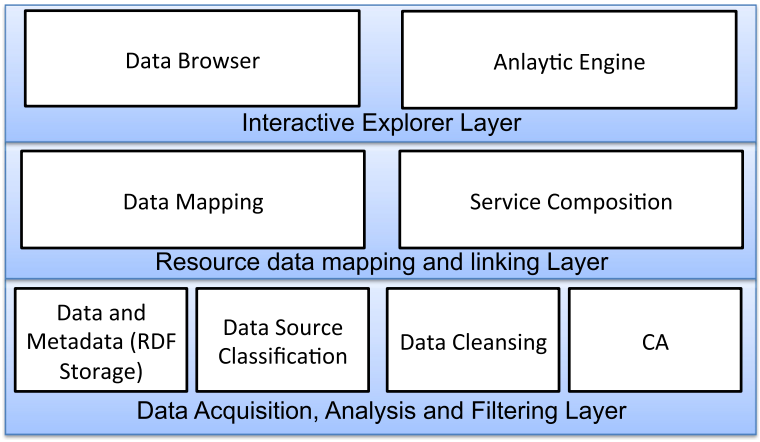

Figure 2 Proposed architectural design.

functionality that we need to meet the overall research objectives. The lowest layer in the architecture consists of distributed and heterogeneous repositories and various sensors that are subscribed to the system. The objective of this layer is data acquisition, cleansing and classification using standard approaches such as APIs or OGC (Open Geospatial Consortium) compliant web services. Existing tools like TheDataTank[a] and CKAN[b] for data access, transformation and publishing (e.g. XML, CSV, JSON or binary structures such as SHP files or relational database) in a RESTful way can be utilised. For data storage Cassandra (un/semi-structured - no SQL), PostgreSQL (relational structured data) and Virtuoso RDF store are selected. However, detailed design and prototype of the bottom two tiers is not within the scope of this paper and is partly covered in [22] and rest is a work in progress.

The resource data mapping and linking layer (middle layer) finds new scenarios and supports workflows to develop relations that were not possible in the isolated data repositories. However it is likely that collected data will be in a number of different formats and semantics due to heterogeneous data sources and hence can benefit from data linking. For example, linked data or open data [23,24] where databases can be browsed to serve queries and find events of interest that were not possible without the availability of linked data. Furthermore, semantic data model can be developed as a layer on top of the linked data to make sense of everything. Once the meta-data of heterogeneous data sources has been populated into meta-data stores, mappings are established between the resources, links are generated and the data is made semantically relevant and browse-able. This data browsing can help end users to select different cross-thematic indicators and variables to perform analytics. Existing metadata formats (such as the European Data Model, Talis Aspire, the Open Library and DBLP as Linked Data) are preferable choices to describe and store meta-data extracted from different sources. The data is then mapped using standardised resource description semantics, e.g. via an RDF store (e.g. Virtuoso DB) which has all the necessary links established between artefacts and resources. In case of linked services, higher level services and mashups can be composed to browse and make use of this data for interesting scenarios. SPARQL, an RDF query language, then can be used to retrieve and manipulate data stored in Resource Description Framework format.

An analytic engine in top layer processes the data for application specific purposes. The engine utilises the data that is available in the linked data layer and helps users in submitting queries, application specific algorithms and workflows to find information from the data repositories. In this respect, Big Data Mining is recently a new trend used to identify large data sets due to complexity, cardinality and continuality [25,26]. Big Data Mining techniques are increasingly becoming an important and effective way in various data driven applications such as network traffic risk analysis, business data analysis etc. These techniques will be extremely useful to generate non-obvious relations and associations from huge data available from public services of smart future cities.

Since the main focus of this paper is smart city data analytics, we'll mainly focus on the analytic engine and explain in detail. For analytic engine, various statistical modelling, machine learning and data mining techniques can be applied. Also, existing tools such as RapidMiner and R in combination with Hadoop MapReduce [8] can be utilised to mine the city data at scale. In literature Big data mining is considered much more and complex than traditional data mining currently in practice [27]. This is true for smart city data analytics because multi-disciplinary nature of city data can help in formulating a variety of city

application scenarios. In this regard, some of the possible components for cloud based big data mining or analysis can be data processing / integration, classification, clustering, data reduction, visualisation, and finding association rules as depicted in Figure 3. It is not necessary to use all these components. Depending upon the application, subsets of components may be needed for data analysis. For example, for the open data use case, algorithms from only two components i.e. data processing and finding association rules are needed. All these components are well known components in data mining [28]. Furthermore, these components can benefit from state-of-the-art tools such as Apache Mahout and R for cluster based scalable machine learning.

A use case: analytics using Bristol open data

This section describes the Bristol Open Data and how it has been used to identify correlations between different urban indicators. Cities usually preserve data about land use, quality of life, health and wellbeing, population, economy and employment, education, transport infrastructure, energy consumption, housing and buildings, local climate or environment (green infrastructure, air quality and noise) etc. With the open data initiative, a lot of such data is made available to citizens and other stakeholders for exploitation. From planning perspective such data may possess information patterns which can be used for predictive analytics for defining new development indicators and simulate future scenarios to support decision making. For instance, Vienna Open Data Portal[c] and Bristol Open Data Portal[d] are two suitable examples which provide open data through OGC web service interfaces such as WFS and WMS (in different formats e.g. JSON, CSV and MS Excel tables). However, cities also possess a huge amount of other data which is not published as open data and hence the available data through open data portals may not comply with all the criteria of Big Data, i.e. volume, velocity, variety and veracity [29].

Considering Bristol open data as an example scenario, the data covers different geospatial scales such as LSOA01 (Lower Layer Super Output Area), Ward, Neighbourhood or city scales. The data is collected from variety of sources including local agencies (e.g. births, deaths, accidents, crimes, energy consumption, air quality, noise etc), Office of National Statistics (e.g. population census) and/or citizens perception (e.g. surveys about quality of life, attractiveness of public spaces), etc. Using the Bristol open dataset a simple analytics scenario can show correspondence between different variables such as health and wellbeing, mortality, air quality, quality of life, house prices, household income and crime events to perform a comparative analysis between different wards in the city to predict and assign priority ranking about more likely liveable area in Bristol in future. Such information can be useful for local stakeholders to get awareness about their locality and local administration to plan appropriate actions to avoid any social economic and digital divide in the city.

An analysis of the Bristol open dataset indicates that there are numerous possibilities for the development of smart solutions. Availability of historical data varies for each indicator for each year e.g. Census and Population data exist for every year between 2001 to 2012 as compared to Quality of Life: crime and safety between years 2005 to 2013. For each indicator there are number of variables or questions i.e. on average approx. 20 and not all these variables have consistent availability of data for all specified years. Mostly data is in statistical aggregated form for different levels of geographical scales. On the one

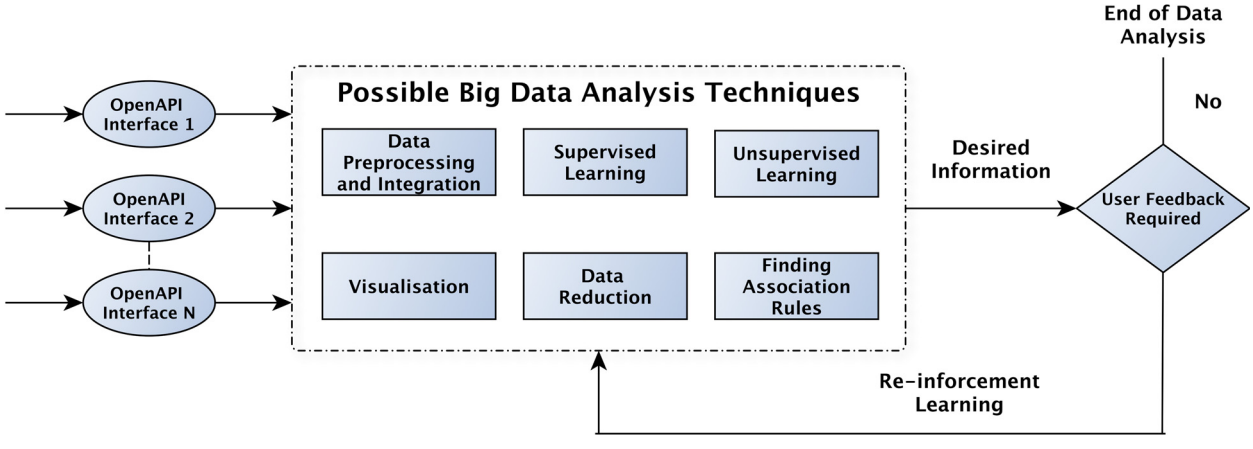

Figure 3 Big data analytic components for analytical engine.

hand the aggregated form of the data helps in avoiding privacy concerns but on the other hand it reduces the overall volume of the data. This situation varies from one city's open data portal to another. Nevertheless, this data can be analysed using appropriate cloud based processing infrastructures to indicate statistical correlations about various indicators e.g. health, employment and citizen perception about selected indicators such as quality of life, as demonstrated in Section "Prototype implementation".

Prototype implementation

For the prototype implementation the Quality of Life (QoL) data was chosen which has a number of indicators to measure the QoL such as Crime and Safety, Culture and Leisure, Economy and Employment etc. Each indicator is measured via a questionnaire designed to obtain citizens' opinion about the relevant indicator for their area. The dataset available contains aggregated values for the questionnaire responses over a number of years as indicated in Figure 4. The size of the data was approximately 0.7 MB.

Prototype application architecture

To demonstrate the real world applicability of the proposed Big data analysis architecture, a basic prototype application (mainly analytic engine) using MapReduce [11] shown in Figure 4 was developed using Hadoop and Spark. The survey questionnaires used by the Bristol City Council give an indication of what the citizens think about the various indicators. However, there is no implicit quantifiable measure of the various indicators. Such a quantifiable measure can be helpful for decision makers when analysing the Quality of Life in Bristol. For example, planners can use the quantitative measures for the indicators spread over years to assess positive or negative trends or effects of certain policies. Moreover, they can also find statistical correlations between the various indicators to determine whether, and to what extent, one indicator affects another. However, since there are a number of questions for each indicator, and each indicator is measured separately for each geographical unit, the size of data is significant. For example, the 2007 questionnaire for assessing the public perception of

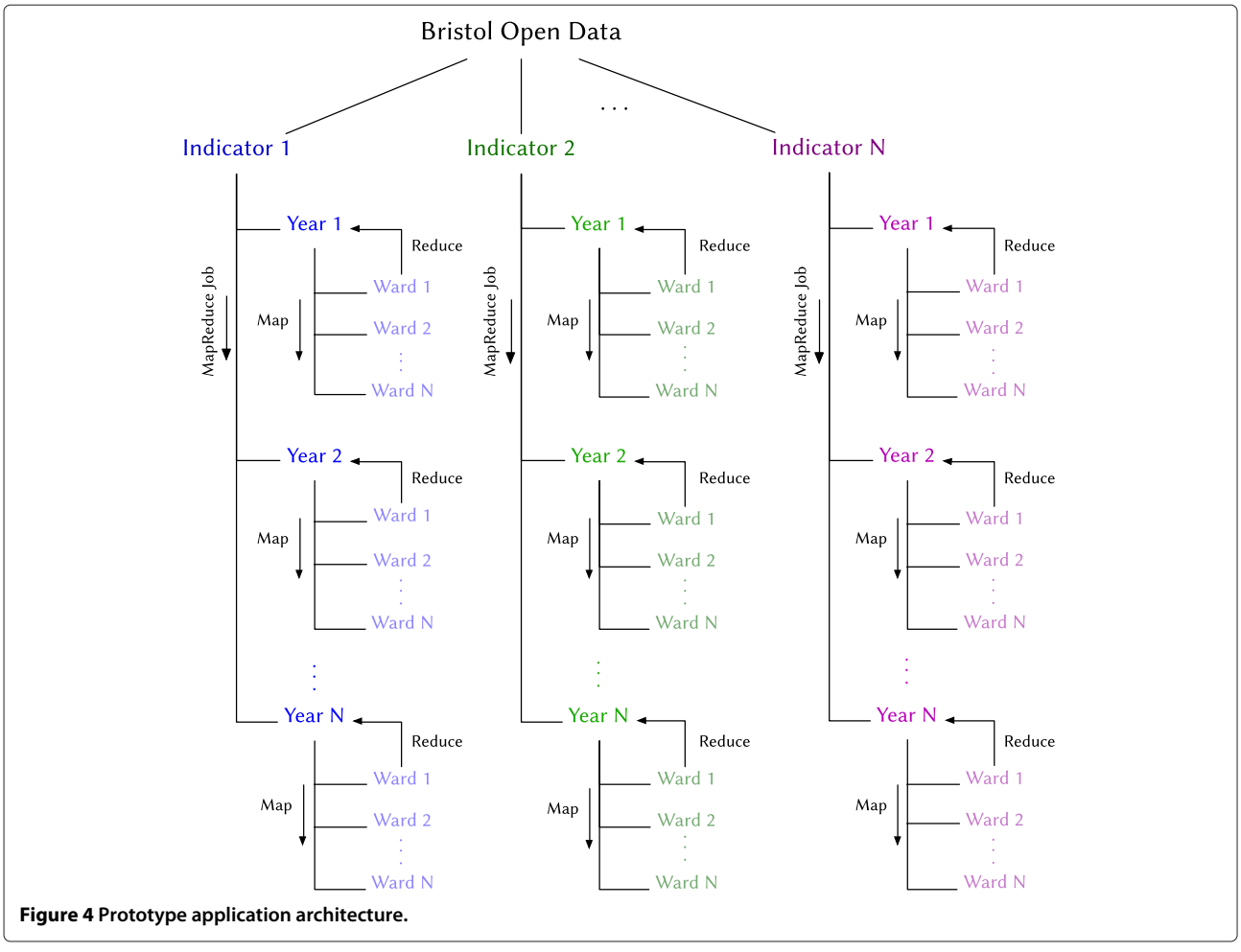

Figure 4 Prototype application architecture.

Crime and Safety consists of approximately 20 questions. Moreover, each question was asked for approximately 40 wards of Bristol which makes 800 questions in total. Since there are responses for 8 years in total for 2 indicators, the total number of questions comes to 12,800 (800x2x8).

To calculate the aforementioned quantitative measures, we propose a hierarchical organisation of the available Bristol open data. For each indicator, we categorised the data by year during preprocessing to facilitate parallelisation by the MapReduce framework. For each year that the data was available, the questionnaire responses were already categorised by wards. In a MapReduce-based application, the data is split automatically by the Hadoop framework and assigned to worker units called Mappers in a step called Mapping. The Mappers perform their programmed tasks, and the data produced as a result is handed over to Reducers after some shuffling and sorting. The Reducers are then responsible for further processing the data into meaningful information. In this setup, application programmers are only concerned with the high-level infrastructure-agnostic architecture of the application.

For this application, the data for each ward was assigned to individual Mappers for processing; calculating the indicator value for that ward for that particular year based on the questionnaire responses. Once the ward-based values were calculated, they were passed to the Reducers. In this step the ward-based values were combined into a single, overall value for that year. In this manner, yearly measures for the various indicators were obtained. Moreover, the values could now be used to establish trends for the various indicators over time. This in turn allowed us to calculate correlations between them. The entire mapping and reducing phase can be considered as data preprocessing in terms of the proposed architecture while the correlation calculation may be considered data mining component.

Experimental setup and results

The experimental setup is shown in Figure 5. For the purposes of this prototype data pertaining to Crime and Safety and Economy and Employment was chosen from the Bristol Open data catalogue. For each indicator, the data was available for 8 years from 2005 to 2012. Each source survey for the data was conducted for 40 wards of Bristol and the questionnaires consisted of up to 20 questions. The data consisted of the number of people that answered yes to the various questions. The cloud infrastructure used consisted of two compute and data nodes with 2GB RAM and single-core processors each. Both nodes were running in virtual machines managed by VMWare Workstation 10. The virtual machines were hosted on a Dell PowerEdge R415 server with an AMD Opteron 4332 HE hexa-core processor and 64GB RAM. The Hadoop infrastructure comprised YARN 2.3.0 along with HDFS 2.3.0. For Spark version 1.1.0 was used.

As mentioned previously the data was parallelised at the ward level. That is to say that the data for each ward was assigned to an individual mapper for calculating the overall index based on the responses. This resulted in a total of 640 mappers divided between two MapReduce jobs. The output of the mappers consisted of measurement indices for the various wards for individual years. The measurement indices were calculated by averaging the number of positive responses to each question for each ward. The reducers were then responsible for aggregating the indices for each year into a yearly index. The final output consisted of two sets of eight values each; one for each year for each indicator. The results of the Spark and Hadoop implementations are discussed subsequently.

Hadoop implementation

For this implementation, a total of 384 mappers and 3 reducers were created. Each job took approximately 15 seconds to execute. The experiments were conducted

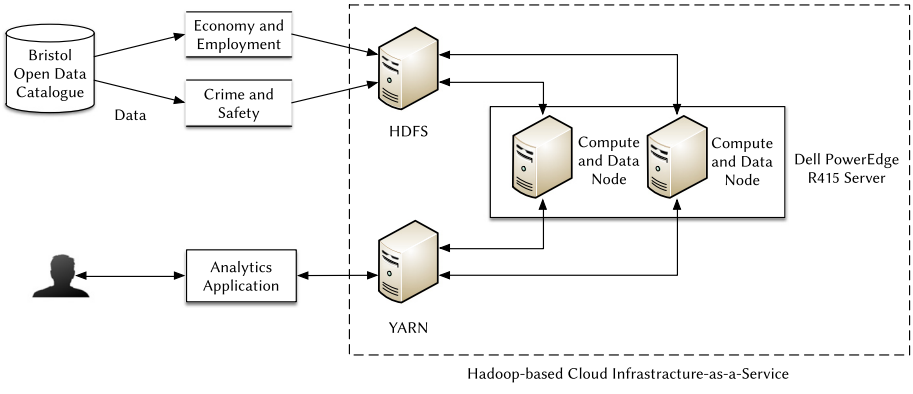

Figure 5 Experimental setup.

Table 1 Hadoop implementation results

Compute nodes	Execution mode	Per task execution time	Total execution time
1	local	~0.01 sec	~5 sec
1	cluster	~15 sec	~25 min
3	cluster	~15 sec	~12 min

using three different configurations. The configurations and their corresponding execution times are shown in Table 1.

In *local* mode, the job is executed on the local machine without contacting the Hadoop ResourceManager. In *cluster* mode, the job is submitted to the ResourceManager which is then responsible for scheduling it to any compute nodes available. It is clear from the resuls that Hadoop incurred a significant overhead when a job was executed on the cluster via the ResourceManager.

Spark implementation

For this implementation a similar setup was used. The configurations used are shown in Table 2.

Compared to Hadoop, Spark incurred significantly less overhead when submitting jobs on the cluster. This is likely due to inexpensive data access operations. Therefore, Spark proved to be more appropriate for the selectd Bristol open dataset.

Application results

This section presents the results of the analyses from an application point-of-view and illustrates how they can be beneficial for urban planners. The indices obtained from the experiments are plotted in Figure 6.

The figure shows the variation in the responses from citizens over the years for the two indicators. The results show that between 2005 and 2012, public approval for economic and employment opportunities declined with an upward turn in 2012. On the other hand public approval of the crime and safety situation improved in 2007 but dipped back down in subsequent years. Since we have such data available for individual wards, we can calculate such trends for each ward as well. It should be noted that the questionnaires evolved over the years so their length and type of questions varied from year to year. Therefore, the overall index is just the average value of the positive responses for every question that was available.

Table 2 Spark implementation results

Compute nodes	Execution mode	Per task execution time	Total execution time
1	local	~0.03 sec	~6 sec
3	cluster	~0.6 sec	~40 sec

The Pearson's correlation calculated between both indicators based on the available data was 0.2305 [30]. This is the most common measure of statistical correlation between two datasets. The value of this measure is always between 1 and -1. 1 indicates a strong correlation, 0 indicates no correlation, and -1 indicates a strong negative correlation. The value 0.2305 indicates that there is a weak correlation between the two indicators in question. The positive value indicates that improvement in the economic and employment opportunities improves the crime and safety situation in the city to some extent. A causal relationship can also be intuitively established between the two indicators. This correlation can be verified visually through Figure 6. Even though the overall trends track each other, there are still instances where the correlation is weak. For example, from 2005 to 2006 there is a sharp decline in the economic and employment opportunities. However, the crime and safety situation improves slightly. Then from 2008 to 2009, the economic and employment opportunities do not vary significantly. However, there is a slight decline in the crime and safety situation. These minor differences in trends lead to an overall weak correlation. However, caution should be exercised when drawing conclusions since the sample set is too small to be statistically valid. Since the purpose of this prototype is to demonstrate how cloud infrastructures can be used to analyse big datasets, a discussion about the scalability of the application is relevant here. It is presented in the next section.

Scaling the application

For demonstration purposes a small sample set is used for this case study. Generally, such data includes hundreds of indicators and are in large quantity. Therefore, a discussion about the scalability of this approach is required. Given the structure of the dataset used, there are several possible expansion scenarios. The following possibilities, in order of the existing hierarchy, exist:

1. The number of indicators might increase. Also, the number of years for which the data was collected might increase.
2. The number of wards might increase (e.g. larger cities). Also, the number of questions asked for each ward might increase.
3. The number of data sources might increase.

Each of these possibilities are discussed below:

If the number of indicators or years increase: the Bristol Open datasets consists of more than 10 indicators in total. Therefore, even though only 2 are used in the experiments shown in this paper, real-world applications will likely consist of more. Moreover, as more and more data is collected, the

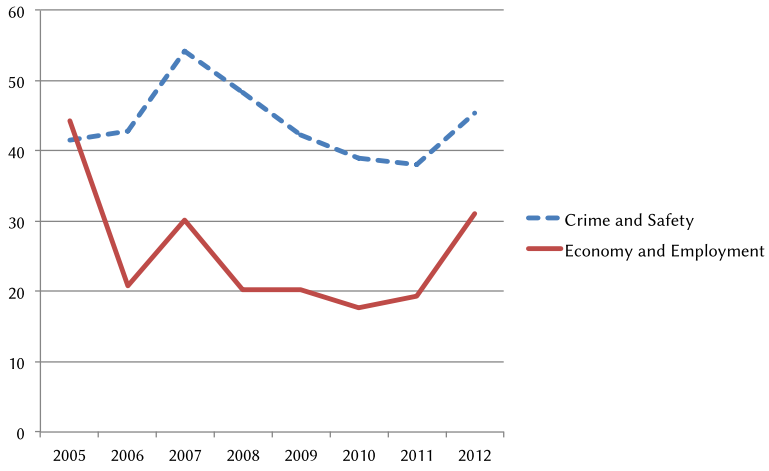

Figure 6 Citizen perception regarding crime and safety vs economy and employment.

number of years it is available for will also increase. If the number of indicators or number of years increase, the only thing required is to create more MapReduce jobs. The application will automatically scale after that. For example, currently for Safety and Crime, data is available for 9 years from 2005 to 2013 and every year this dataset will grow.

If the number of wards or number of questions asked for each ward increase: since the data for each ward is assigned to individual Mappers, increasing the number of wards would also mean an increase in the number of Mappers that are created. Once again, this will be handled automatically by the application. Therefore, the application will also scale automatically and no change would be required. This is possible, for example if data for larger cities such as London is being analysed, which has 624 wards as opposed to Bristol's 40. Also, for city-region analysis where data from wards in a city and metropolitan area need to be analysed e.g. urban sprawl and transport (e.g. daily commuters) correlations to determine environmental impact assessment e.g. CO2 emissions.

If the number of data sources increase: additional data sources might be considered for calculating the indices for the various indicators. Such additional data sources can be UK's Office for National Statistics, OpenStreetMap[e] or Open Data Portal[f]. Moreover, the Bristol Open Data Portal contains statistics of responses to questionnaire-based surveys conducted by the Bristol City Council. It is these statistics that have been used in this experiment. However, one may also wish to compare the perception of crime to the reality. For this purpose the crime statistics dataset available from the Police UK [g] website would be required which is 4.51 GB. Another possibility could be adding semantic information indicating the relative weight of each question as well as additional factors that contribute towards calculating the index. In both these cases, the nature of the prototype application does not need to be changed. However, this might not be true for all kinds of data sources.

The next section concludes this paper.

Conclusions and future directions

Smart cities provide an opportunity to connect people and places using innovative technologies that helps in better city planning and management. At the core of smart cities are the collection, management, analysis and visualisation of huge amount of data that is generated every minute in an urban environment due to socioeconomic, anthropogenic or natural environmental events or other activities. Smart cities data can be collected directly from variety of sensors, smart phones, citizens and integrated (or linked) with city data repositories to perform analytical reasoning and generate required information (e.g. for end users) or new knowledge for decision-making for better urban governance. Innovations in information and communication technological provide the opportunity to manage and process smart city data and provide timely and necessary information to relevant stakeholders for decision making.

In this paper we discussed the cloud based big data analytics for smart future cities. Several considerations need to be carefully planned such as data collection, preparation, semantic linking and use of appropriate data mining, machine learning or statistical analytical techniques. In addition, due to multidisciplinary nature of smart city application domains, engagement with domain experts is needed to identify basic relationships and dependencies

between different data elements. The proposed architecture provides basic components to build necessary functionality for a cloud based big data analytical service for smart cities data. As a proof of concept, we have developed a prototype using MapReduce that demonstrates how cloud infrastructure can be used to analyse a sample set of Bristol Open Data. The protype has been implemented using Hadoop and Spark and the results are compared. The results show that Hadoop incurs significant overhead when jobs are submited to the cluster, likely due to expensive data access operations. Comparatively, Spark is much faster and incurs significantly less overhead. Therefore, Spark is more appropriate for the chosen Bristol open dataset.

Technically the dataset accessible through the Bristol Open Data portal does not fully constitute Big Data due to its aggregated form. However, the proof-of-concept shows how such computing infrastructures can be applied to Big Data solutions. Based on the experiment results, we discussed the suitability of elastic nature of the cloud resources to fulfil the demand of smart cities data analytic needs. The prototype implementation indicates usefulness of cloud based infrastructure for smart city data analytics.

In prototype application, the reason of using open data was that most of city administration data is not available in public domain. On the one hand some cities aggregate daily data into months (even years) and publish, that hugely reduces the overall size/volume of such data. The overall dimensions are high but such aggregated data lack in getting more specific geo-coordinates/locations. This results in privacy protection and easy management of such data but provides a very small sample set to get more detailed insights and identify more precise correlations between data elements. Nevertheless, over the period of time such data tends to grow but may not be as detailed and big as is in the case of other big cities. Bristol's open data portal has such aggregated data sets which fall in few hundred K bytes to Mega bytes.

In contrast to above aggregated data publishing, some cities tend to share more detailed data with frequent updates (i.e. velocity) on daily basis. Such detailed data is large in size as well as in dimensions (i.e. variety). It can be used to get more detailed insights to derive predictive analysis. For instance, San Francisco's open data portal[h] has such open data sets with records in millions and data size reaching from few hundred Mega bytes to 10s of Giga bytes. For example, the crime incidents dataset alone consists of 800,000+ records starting from early 2003 and is updated daily. The size of the dataset is already 300+ MB as of the revising of this paper. The approaches discussed in this paper can be applied to such larger data by grouping yearly datasets.

Our future research work is to scale the technical infrastructure to identify correlations between other indicators in available open data and investigate how semantic sources such as RDF stores can be utilised in Open datasets. The aim of this endeavour would be to identify technical implications and limitations and suggest viable solutions.

Endnotes

[a]The Data Tank: Accessible from URL: https://github.com/tdt/ [Last Accessed: 28 January 2015].

[b]CKAN open-source data portal platform: Accessible URL: http://ckan.org/ [Last Accessed: 28 January 2015].

[c]Vienna Open Data Catalogue: Accessible from URL: https://open.wien.at/site/datenkatalog/ [Last Accessed: 28 January 2015].

[d] Bristol Open Data portal: Accessible from URL: http://profiles.bristol.gov.uk/ [Last Accessed: 28 January 2015].

[e]OpenStreetMap: Accessible from URL: http://www.openstreetmap.org/#map=5/54.910/-3.432 [Last Accessed: 28 January 2015].

[f]UK Open Data Portal: Accessible from URL: http://data.gov.uk/ [Last Accessed: 28 January 2015].

[g]Crime and Policing England, Wales and NI, Crime Data: Accessible from URL: http://www.police.uk [Last Accessed: 28 January 2015].

[h]San Francisco Open Data portal: Accessible from URL: http://data.sfgov.org/ [Last Accessed: 28 January 2015].

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
ZK carried out the research in identifying the scope of this research. ZK provided smart cities context, developed use case and modelled design and development process for cloud based Big data analytic service. AA developed the proposed architecture and with the help of MAT identified existing technologies to be applied in different architectural components of cloud based Big data analytic service. MAT and KS provided inputs based on data mining and analytic techniques for the big data analytic service. KS also designed and implemented the prototype application and also critically discussed its results and scalability aspects. All authors read and approved the final manuscript.

Acknowledgements
We acknowledge early discussion and feedback on architectural components from Dr. Saad Liaquat of the University of the West of England, Bristol, UK.

Author details
[1]Faculty of Environment and Technology, Department of Computer Science and Creative Technologies, University of the West of England, Bristol, UK. [2]Faculty of Business, Computing and Law, School of Computing and Mathematics, University of Derby, Derby, UK. [3]School of Computer Science and Digital Technologies, University of Northumbria, NE1 8ST, Newcastle upon Tyne, United Kingdom.

Received: 1 September 2014 Accepted: 7 January 2015
Published online: 18 February 2015

References

1. Bandyopadhyay D, Sen J (2011) Internet of things: Applications and challenges in technology and standardization. Wirel. Pers. Commun. 58(1):49–69
2. Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of things (iot): A vision, architectural elements, and future directions. Future Gener. Comput. Syst. 29(7):1645–1660
3. Ludlow D, Khan Z (2012) Participatory democracy and the governance of smart cities. In: Proceedings of the 26th Annual AESOP Congress, Ankara, Turkey
4. Khan Z, Kiani SL (2012) A cloud-based architecture for citizen services in smart cities. In: ITAAC Workshop 2012. IEEE Fifth International Conference on Utility and Cloud Computing (UCC), Chicago, IL, USA. pp 315–320. IEEE
5. Suciu G, Vulpe A, Halunga S, Fratu O, Todoran G, Suciu V (2013) Smart cities built on resilient cloud computing and secure internet of things. In: 19th International Conference on Control Systems and Computer Science (CSCS), Bucharest, Romania. pp 513–518
6. Ferguson M Architecting A Big Data Platform for Analytics. White paper. Available online: http://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14333usen/IML14333USEN.PDF, [Last accessed: 21st January, 2015]
7. Hurwitz J, Nugent A, Halper F, Kaufman M (2013) Big Data for Dummies. 1st edn. John Wiley & Sons, Inc., Hoboken, New Jersey
8. Lu S, Li MR, Tjhi CW, Leen KK, Wang L, Li X, Ma D (2011) A framework for cloud-based large-scale data analytics and visualization: Case study on multiscale climate data. In: Proceedings of the 3rd IEEE International Conference on Cloud Computing Technology and Science, Nov 29-Dec 1 2011, Divani Caravel, Athens, Greece. pp 618–622
9. Burnap P, Rana O, Williams M, Housley W, Edwards A, Morgan J, Sloan L, Conejero J COSMOS: Towards an Integrated and Scalable Service for Analysing Social Media on Demand. International Journal of Parallel, Emergent and Distributed Systems. http://orca.cf.ac.uk/59478/
10. Ahuja PS, Moore B (2013) State of big data analysis in the cloud. Network and Communication Technologies 2(1):62–68
11. Qin X, Wang H, Li F, Zhou B, Cao Y, Li C, Chen H, Zhou X, Du X, Wang S (2012) Beyond simple integration of rdbms and mapreduce - paving the way toward a unified system for big data analytics: Vision and progress, pp.716–725. Second International Conference on Cloud and Green Computing, Xiangtan, China
12. Moraru A, Mladenic D (2012) Complex event processing and data mining for smart cities. In: Conference on Data Mining and Data Warehouses (SkiDD 2013), Held at the 15th International Multiconference on Information Society (IS-2012), 8th October 2012, Ljubljana, Slovenia
13. Mutizwa-Mangiza ND, Arimah BC, Jensen I, Yemeru EA, Kinyanjui MK (2011) Cities and climate change: Global report on human settlements. In: Global Report on Human Settlements, p.250. UN-HABITAT
14. European Environment Agency (2006) Urban sprawl in Europe - the ignored challenge. Technical report, European Commission. ISBN: 92-9167-887-2, Available from: http://www.eea.europa.eu/publications/eea_report_2006_10/eea_report_10_2006.pdf [Last accessed: 21st January, 2015]
15. Khan Z, Ludlow D, McClatchey R, Anjum A (2012) An architecture for integrated intelligence in urban management using cloud computing. J Cloud Comput Appl Adv Syst Appl 1(1):1–14. http://www.journalofcloudcomputing.com/content/1/1/1
16. Khan Z, Pervez Z, Ghafoor A (2014) Towards cloud based smart cities data security and privacy management. In: 2014 7th IEEE/ACM International Conference on Utility and Cloud Computing - SCCTSA Workshop, 8th–11th December, London, UK. pp 806–811
17. IBM, Zikopoulos P, Eaton C (2011) Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media. http://freecomputerbooks.com/Understanding-Big-Data.html
18. Naphade M, Banavar G, Harrison C, Paraszczak J, Morris R (2011) Smarter cities and their innovation challenges. Computer 44(6):32–39
19. Open Government Data. http://opengovernmentdata.org/ [Last accessed: 21st January, 2015]
20. da Silva WM, Alvaro A, Tomas GHRP, Afonso RA, Dias KL, Garcia VC (2013) Smart cities software architectures: a survey. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing, New York, NY, USA, pp. 1722–1727. http://jwcn.eurasipjournals.com/content/2012/1/247
21. Nathalie M, Symeon P, Antonio P, Kishor T (2012) Combining cloud and sensors in a smart city environment. EURASIP journal on Wireless Communications and Networking 247:1–10
22. Khan Z, Kiani SL, Soomro K (2014) A framework for cloud-based context-aware information services for citizens in smart cities. Journal of Cloud Computing Applications: Advances, Systems and Applications 3(14):1–17
23. Linked Data - Design Issues. http://www.w3.org/DesignIssues/LinkedData.html [Last accessed: 21st January, 2015]
24. Linked Open Data: The Essentials - A Quick Start Guide for Decision Makers. http://www.semantic-web.at/LOD-TheEssentials.pdf [Last accessed: 21st January, 2015]
25. Wu X, Zhu X, Wu G-Q, Ding W (2014) Data mining with big data. Transactions on Knowledge and Data Engineering: IEEE 26(1):97–107
26. Fan W, Bifet A (2012) Mining big data : current status and forecast to the future. SIGKDD Explorations: ACM 14:1–5
27. Lin J, Ryaboy D (2012) Scaling big data mining infrastructure: The twitter experience. SIGKDD Explorations: ACM 14:6–19
28. Bishop CM (2007) Pattern recognition and machine learning. Springer. http://www.springer.com/computer/image+processing/book/978-0-387-31073-2
29. Siewert SB (2013) Big data in the cloud: Data velocity, volume, variety, veracity. IBM-Developer Works. Available from: http://www.ibm.com/developerworks/library/bd-bigdatacloud/bd-bigdatacloud-pdf.pdf [Last accessed: 21st January, 2015]
30. Ahlgren P, Jarneving B, Rousseau R (2003) Requirements for a cocitation similarity measure, with special reference to pearson's correlation coefficient. Journal of the American Society for Information Science and Technology 54(6):550–560